

A Bayesian Antedependence Model for Whole Genome Prediction

Wenzhao Yang and Robert J. Tempelman¹

Department of Animal Science, Michigan State University, East Lansing, Michigan 48824-1225

ABSTRACT Hierarchical mixed effects models have been demonstrated to be powerful for predicting genomic merit of livestock and plants, on the basis of high-density single-nucleotide polymorphism (SNP) marker panels, and their use is being increasingly advocated for genomic predictions in human health. Two particularly popular approaches, labeled BayesA and BayesB, are based on specifying all SNP-associated effects to be independent of each other. BayesB extends BayesA by allowing a large proportion of SNP markers to be associated with null effects. We further extend these two models to specify SNP effects as being spatially correlated due to the chromosomally proximal effects of causal variants. These two models, that we respectively dub as ante-BayesA and ante-BayesB, are based on a first-order nonstationary antedependence specification between SNP effects. In a simulation study involving 20 replicate data sets, each analyzed at six different SNP marker densities with average LD levels ranging from $r^2 = 0.15$ to 0.31 , the antedependence methods had significantly ($P < 0.01$) higher accuracies than their corresponding classical counterparts at higher LD levels ($r^2 > 0.24$) with differences exceeding 3%. A cross-validation study was also conducted on the heterogeneous stock mice data resource (<http://mus.well.ox.ac.uk/mouse/HS/>) using 6-week body weights as the phenotype. The antedependence methods increased cross-validation prediction accuracies by up to 3.6% compared to their classical counterparts ($P < 0.001$). Finally, we applied our method to other benchmark data sets and demonstrated that the antedependence methods were more accurate than their classical counterparts for genomic predictions, even for individuals several generations beyond the training data.

WHOLE genome prediction (WGP) using commercially available medium to high density (>50,000) single-nucleotide polymorphism (SNP) panels have transformed livestock and plant breeding. Typically, the allelic substitution effects of all SNP markers are jointly estimated in WGP evaluation models assuming additive inheritance and summed to predict breeding values of each individual animal on the basis of its SNP genotypes (Meuwissen *et al.* 2001). This technology will not only lead to dramatically increased rates of genetic improvement for economically important traits such as meat and milk production in livestock (Wiggans *et al.* 2011) or crop production (Lorenz *et al.* 2011), but would also improve predictions of genetic predisposition to human diseases for personalized medicine (de los Campos *et al.* 2010).

Currently, the number of available SNP markers (m) is typically much greater than the number of animals having

phenotypic records (n). Hence, hierarchical mixed-model or Bayesian approaches have been generally adopted in WGP to efficiently borrow information across these many markers by specifying their corresponding effects to be random. Following Meuwissen *et al.* (2001), these effects are typically specified to be either Gaussian or Student *t*-distributed (BayesA), or a mixture of either of these two densities with a point mass on zero (BayesB). When these effects are specified to be Gaussian, then best linear unbiased prediction of these effects is typically pursued because of computational tractability (VanRaden 2008; Hayes *et al.* 2009); applied to WGP, this procedure is often known as GBLUP (genomic-best linear unbiased prediction). Thus far, the distributional specifications for these various hierarchical modeling approaches have been based on a prior assumption of independence between all such effects.

Gianola *et al.* (2003) anticipated that some of these effects might be spatially correlated within chromosomes such that greater inference efficiency might be provided by modeling these effects as correlated. Their proposed specifications required equally spaced markers and/or within-chromosome correlations depending strictly on physical/linkage map distance between markers. However, the equally spaced assumption is

Copyright © 2012 by the Genetics Society of America
doi: 10.1534/genetics.111.131540

Manuscript received July 18, 2011; accepted for publication November 20, 2011

Supporting information is available online at <http://www.genetics.org/content/suppl/2011/12/01/genetics.111.131540.DC1>.

¹Corresponding author: Department of Animal Science, Michigan State University, 1205 Anthony Hall, East Lansing, MI 48824-1225. E-mail: tempelma@msu.edu

rather tenuous for most currently available SNP marker panels. Even more importantly, the inferred correlation structure is likely to be nonstationary given that it should be primarily driven by the proximity of SNP markers to quantitative trait loci (QTL) of major effects. In other words, we anticipate that the correlation between the inferred effects of adjacent SNPs distal to major QTL would be substantially smaller than those proximal to these QTL.

Antedependence models have been increasingly advocated for the analysis of repeated measures data (Zimmerman and Nunez-Anton 2010) to parsimoniously account for nonstationary correlations between repeated measurements over time. In this article, we develop first-order antedependence counterparts to BayesA and BayesB. Through simulation, a cross-validation study involving the publicly available heterogeneous stock mice data (Valdar *et al.* 2006a,b) and journal-provided reference data (Hickey and Gorjanc 2012) to potentially benchmark our proposed methods with others, we demonstrate that, compared to their conventional counterparts, these antedependence-based WGP models improve the accuracy of genomic merit as well as potentially increase the sensitivity of QTL detection, which is the key objective of genome-wide association studies (GWAS).

Materials and Methods

Conventional WGP model

The base linear mixed model used for WGP is generally written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \mathbf{W}\mathbf{u} + \mathbf{e}. \quad (1)$$

Here, $\mathbf{y} = \{y_i\}_{i=1}^n$ is a $n \times 1$ vector of phenotypes, $\boldsymbol{\beta}$ is an $p \times 1$ unknown vector of fixed effects connected to \mathbf{y} via a known $n \times p$ incidence or covariate matrix \mathbf{X} (e.g., environmental effects) $\mathbf{g} = \{g_j\}_{j=1}^m$ is an $m \times 1$ vector of random SNP effects connected to \mathbf{y} via a known $n \times m$ matrix of SNP genotypes coded as 0, 1, or 2 copies of the minor allele for each SNP (column) and animal (row) in \mathbf{Z} . Furthermore, $\mathbf{u} = \{u_k\}_{k=1}^q$ is a $q \times 1$ vector of random polygenic effects connected to \mathbf{y} via a known $n \times q$ incidence matrix \mathbf{W} , and $\mathbf{e} = \{e_i\}_{i=1}^n$ is the residual vector. We assume that $\mathbf{u} \sim N(0, \mathbf{A}\sigma_u^2)$, where \mathbf{A} denotes the pedigree-derived numerator relationship matrix (Henderson 1976) and is often included in WGP models due to insufficient genome coverage by \mathbf{Z} (Calus and Veerkamp 2007). Furthermore, we specify $\mathbf{g} \sim N(0, \mathbf{G})$, where $\mathbf{G} = \text{diag}(\sigma_{g_j}^2)$ and $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$. From a Bayesian perspective, a subjective prior may be also specified on $\boldsymbol{\beta}$ using $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \mathbf{V}_\beta)$ with $\boldsymbol{\beta}_0$ and \mathbf{V}_β taken as known (Sorensen and Gianola 2002).

Now the distinction between GBLUP, BayesA, and BayesB in Meuwissen *et al.* (2001) depends upon the characterization of \mathbf{G} . If $\mathbf{G} = \mathbf{I}\sigma_g^2$ (i.e., $\sigma_{g_j}^2 = \sigma_g^2 \forall j$), then the model is defined to be GBLUP. If, instead, the diagonal elements of \mathbf{G} are independent random draws from a scaled inverted chi-square distribution, i.e., $\sigma_{g_j}^2 \sim \chi^{-2}(v_g, v_g s_g^2)$ such that $E(\sigma_{g_j}^2) = v_g s_g^2 / (v_g - 2)$, then the model is said to be BayesA

such that marginally g_j is a random draw from a Student t distribution with mean 0, degrees of freedom v_g , and scale parameter s_g^2 (de los Campos *et al.* 2009; Gianola *et al.* 2009). Now BayesB further extends BayesA by including a two-component mixture with one component being $\chi^{-2}(v_g, v_g s_g^2)$ and the other component being a spike or point mass at 0; i.e.,

$$\sigma_{g_j}^2 | v_g, s_g^2 \begin{cases} = 0 & \text{with probability } \pi_g \\ \sim \chi^{-2}(v_g, v_g s_g^2) & \text{with probability } (1 - \pi_g). \end{cases} \quad (2)$$

That is, π_g ($0 < \pi_g < 1$) represents the proportion of SNP markers having no associated genetic effects on the trait of interest.

Clear warnings have been provided on how sensitive inferences using BayesA or BayesB may be to specification of the hyperparameters (de los Campos *et al.* 2009; Gianola *et al.* 2009). It has not been widely appreciated that v_g and s_g^2 are estimable; this recognition is critical as both hyperparameters help define the genetic architecture in BayesA and BayesB. That is, v_g characterizes the variability of $\sigma_{g_j}^2$ about a typical variance component of s_g^2 . Details on how to estimate v_g and s_g^2 in the context of BayesA were previously provided by Yi and Xu (2008). Furthermore, π_g is estimable in BayesB. For both BayesA and BayesB, we specify the prior distribution $v_g \sim p(v_g) \propto (v_g + 1)^{-2}$, similar to what we have previously adopted in other applications (Kizilkaya and Tempelman 2005; Bello *et al.* 2010). Furthermore, we specify $\pi_g \sim p(\pi_g | \alpha_\pi, \beta_\pi) = \text{Beta}(\alpha_\pi, \beta_\pi)$ for BayesB, with values of α_π and β_π chosen to reflect prior uncertainty on π_g . We also specify a proper conjugate prior on s_g^2 in BayesB; i.e., $s_g^2 \sim p(s_g^2 | \alpha_g, \beta_g) = \text{Gamma}(\alpha_g, \beta_g)$ recognizing that the specifications on α_g and β_g become increasingly influential as $\pi_g \rightarrow 1$.

Finally, we specify noninformative priors $\sigma_e^2 \sim \chi^{-2}(-1, 0)$ and $\sigma_u^2 \sim \chi^{-2}(-1, 0)$, which are congruent with specifying uniform priors on σ_e and σ_u , respectively, and in line with recommendations for variance components by Gelman (2006). We similarly and confidently specify $s_g^2 \sim \chi^{-2}(-1, 0)$ in BayesA, given that m is generally large enough for stable inference on s_g^2 without the need for more informative priors.

Antedependence extensions of WGP models

We propose a nonstationary first-order antedependence correlation structure for \mathbf{g} based on the relative physical location of SNP markers along the chromosome(s):

$$g_j = \begin{cases} \delta_1 & \text{if } j = 1 \\ t_{j,j-1} g_{j-1} + \delta_j & \text{if } 2 \leq j \leq m. \end{cases} \quad (3)$$

Here $\delta_j \sim \text{NID}(0, \sigma_{\delta_j}^2)$, $j = 1, \dots, m$, whereas $t_{j,j-1}$ is the marker interval-specific antedependence parameter (Zimmerman and Nunez-Anton 2010) of g_j on g_{j-1} in the specified order. We can rewrite the recursive expression in (3) in matrix notation,

$$\mathbf{g} = \mathbf{T}\mathbf{g} + \delta, \quad (4)$$

where $\delta = \{\delta_j\}_{j=1}^m = (\mathbf{I} - \mathbf{T})\mathbf{g}$ for \mathbf{I} being a $m \times m$ identity matrix, and \mathbf{T} having all null values except for elements $t_{j,j-1}$ at the corresponding subscript addresses. It can be readily seen using Equation 4 that $\text{var}(\mathbf{g}) = \mathbf{G} = (\mathbf{I} - \mathbf{T})^{-1}\Delta(\mathbf{I} - \mathbf{T})^{-1}$, where $(\mathbf{I} - \mathbf{T})^{-1}$ is a lower triangular matrix with diagonal elements equal to 1 and $\Delta = \text{diag}\{\sigma_{\delta_j}^2\}_{j=1}^m$. As further illustrated in supporting information, File S1, from the $\mathbf{G}^{-1} = (\mathbf{I} - \mathbf{T})'\Delta^{-1}(\mathbf{I} - \mathbf{T})$ is a readily determined tridiagonal matrix (Zimmerman and Nunez-Anton 2010), which is important as it facilitates inference on \mathbf{g} .

Some of the other developments closely follow the BayesA and BayesB models of Meuwissen *et al.* (2001). That is, we specify $\sigma_{\delta_j}^2 \sim \chi^{-2}(v_{\delta}, v_{\delta}s_{\delta}^2)$ in a model that we label *ante-BayesA*. Similarly, we propose an *ante-BayesB* model whereby we specify a mixture similar to Equation 2 except that it is specified on $\sigma_{\delta_j}^2$; *i.e.*, a mixture of point mass on zero with probability π_{δ} and scaled inverted chi-square prior $\chi^{-2}(v_{\delta}, v_{\delta}s_{\delta}^2)$ with probability $(1 - \pi_{\delta})$. As we suggested earlier for π_g , we believe that π_{δ} is estimable such that *ante-BayesA* is merely a special case of *ante-BayesB*. In turn, *BayesA* is merely a special case of *ante-BayesA*, as is *BayesB* of *ante-BayesB*, when $\mathbf{T} = \mathbf{0}$; *i.e.*, $t_{j,j-1} = 0 \forall j$.

These antedependence extensions, nevertheless, do require inference on the unknown $m - 1$ nonzero elements $\{t_{j,j-1}\}_{j=2}^m$ of \mathbf{T} . Borrowing from Daniels and Pourahmadi (2002) and Bello *et al.* (2010), we specify $t_{j,j-1} \sim N(\mu_t, \sigma_t^2)$ as a conjugate prior in both *ante-BayesA* and *ante-BayesB*, thereby allowing flexible inference on the nonstationary correlation structure in \mathbf{G} . However, it should be further noted that if interval $j, j - 1$ specifies that between the last SNP of one particular linkage group or chromosome and the first SNP in the arbitrarily subsequent linkage group, then we set the corresponding $t_{j,j-1} = 0$. The remaining prior specifications are specified on the hyperparameters that essentially characterize the hypothesized genetic architecture of the trait and are virtually identical to those previously prescribed for *BayesA* and *BayesB*; *i.e.*, $v_{\delta} \propto (v_{\delta} + 1)^{-2}$, $s_{\delta}^2 \sim \text{Gamma}(\alpha_s, \beta_s)$, $\pi_{\delta} \sim \text{Beta}(\alpha_{\pi}, \beta_{\pi})$ with $\alpha_{\pi}, \beta_{\pi}$, α_s , and β_s again all specified as known. Similarly, we also estimate μ_t and σ_t^2 by placing subjective priors, $\mu_t \sim N(\mu_{t0}, s_{t0}^2)$ and $\sigma_t^2 \sim \chi^{-2}(v_t, v_t s_t^2)$ on these key hyperparameters, where μ_{t0}, s_{t0}^2, v_t and s_t^2 are specified to be known.

As in Meuwissen *et al.* (2001) and subsequent work, our implementation strategy is based on the use of Markov chain Monte Carlo (MCMC) methods; however, we also additionally infer key hyperparameters, *i.e.*, $v_g(v_{\delta}), s_g^2(s_{\delta}^2)$, and $\pi_g(\pi_{\delta})$, that characterize the genetic architecture of the trait, as alluded to earlier. Further details on the full conditional densities and any necessary Metropolis–Hastings strategies to sample from the joint posterior density of all unknown parameters using MCMC are provided in File S1.

Simulation study

We compare the performance of *BayesA* and *BayesB* with their antedependence counterparts, *ante-BayesA* and *ante-*

BayesB, in a simulation study. Twenty replicated data sets were each generated from a base population containing 50 unrelated males and 50 unrelated females. Each data set underwent random mating while maintaining constant population size for 6001 generations beyond the base population. The entire genome was composed of one chromosome of length 1 M. All of 20,001 potential SNP markers were equally spaced on this genome with a potential QTL placed directly in the middle of each interval of adjacent markers. In the base population, all 20,000 QTL and 20,001 SNP marker alleles were coded as monomorphic. The number of simulated crossover events per meiosis was generated from a Poisson (mean 1) distribution with the location of the crossover events uniformly distributed throughout the chromosome in accordance with the Haldane mapping function. The mutation rate for both QTL and SNP markers was specified to be 10^{-4} per locus per generation and to be recurrent, that is, switching between one of two alternative allelic states 0 and 1 whenever mutation occurred so as to ensure biallelic loci (*e.g.*, Coster *et al.* 2010; Daetwyler *et al.* 2010).

In Generation 6001, all SNP markers and QTL with a minor allele frequency (MAF) < 0.05 were discarded. We then randomly selected only 30 of the remaining QTL and their corresponding allelic substitution effects. For each of these $k = 1, 2, \dots, 30$ QTL, an allelic substitution effect (α_k) was drawn from a reflected gamma distribution with shape parameter 0.4 and scale parameter 1.66, with a positive or negative sign on α_k sampled with equal probability. The genetic variance at QTL k was determined to be $2p_k(1-p_k)\alpha_k^2$, where p_k is the MAF at QTL k . The total genetic variance was subsequently determined to be the summation of these terms across the 30 selected QTL, *i.e.*, as $2\sum_{k=1}^{30} p_k(1-p_k)\alpha_k^2$. Now the true breeding values (TBV) were defined to be a genotype-based linear function of the 30 generated QTL effects, which, because these QTL were located between various SNP, are not subsets of \mathbf{g} . These TBV were further scaled such that the total genetic variance was 1 as per Meuwissen and Goddard (2010). Residual effects were, in turn, sampled from a standard normal distribution, such that the heritability was 0.50. That is, each phenotypic record was generated by adding the TBV for that animal plus its corresponding residual. Hence, 100 animals with known phenotypes and genotypes in Generation 6001 were simulated for inferring upon the SNP effects, using each of the competing methods. Genotypes and the TBV for each of 100 offspring were also generated in Generation 6002, on the basis of randomly mating animals in Generation 6001.

For each of the 20 replicated data sets, the effect of six different marker densities on the comparison between the competing methods were investigated by selecting every 1, 4, 7, 10, 15, and 20 SNP markers from those with MAF > 0.05 . That is, the data sets were used as a blocking factor in comparing different marker densities for the accuracy of predicting genetic merit in Generation 6002, using each of the four different methods: *BayesA*, *BayesB*, *ante-BayesA*, and *ante-BayesB*. Accuracy was defined as the correlation

between estimated breeding values (EBV) for Generation 6002, using just Generation 6001 phenotypes and genotypes data, and the corresponding TBV of Generation 6002. These EBV are based on the posterior mean ($\bar{\mathbf{g}}$) of \mathbf{g} ; *i.e.*, EBV are elements of $\mathbf{Z}\bar{\mathbf{g}}$. Comparisons between the BayesA/BayesB procedures and their antedependence counterparts were also drawn for inference on the key hyperparameters that characterize genetic architecture. This was conducted using multifactorial ANOVA on the posterior means using replicate as the blocking factor for assessing the importance of model and marker density and their interaction across the 20 replicates. Furthermore, an assessment of the relative ability of ante-BayesB compared to BayesB to identify the top QTL by genetic variance was based on the difference in posterior probabilities of δ_j and g_j , respectively, of adjacent SNP markers being nonzero. As QTL were placed between SNP markers and never on top of SNP markers, we calculated this *probability of association* by determining the proportion of MCMC cycles that either or both of the two markers adjacent to the known QTL were chosen to be nonzero within each analysis.

All comparisons were based on the linear mixed model in Equation 2 with \mathbf{X} being a column vector of ones, except that polygenic effects (\mathbf{u}) were ignored for simplicity and computational tractability.

Application to heterogeneous stock mice data set

We used a data set publicly available from the Wellcome Trust (<http://gscan.well.ox.ac.uk/>), which includes phenotypic records on 2296 mice, each genotyped for 12,147 SNP markers. This data resource, which also includes pedigree information, was based on an advanced intercross mating among eight inbred strains after 50 generations of random mating (Valdar *et al.* 2006a). The average linkage disequilibrium (LD), as measured by r^2 between adjacent markers is 0.62 (Legarra *et al.* 2008), which is high compared to commonly used SNP panels available for livestock populations. For example, the average r^2 between adjacent markers in most commercially available livestock SNP panels ranges from 0.10 to 0.37 for markers that are generally around 100 kb apart (Du *et al.* 2007; de Roos *et al.* 2008; Abasht *et al.* 2009; Bohmanova *et al.* 2010).

Given this high pairwise LD, we considered only a random subset of all markers from this data set to ensure adjacent marker LD levels that are representative of livestock populations. We first excluded SNP markers if the percentage of missing genotypes across samples was $>10\%$ or if the MAF was $<2.5\%$. We also discarded animals having greater than 20% missing SNP genotypes. We then randomly selected 50 SNP markers from each of the 19 autosomes, leading to an average LD of r^2 of 0.35 between adjacent markers. The resulting data set then involved records on 1917 animals with genotypes on 950 SNPs. The phenotypes and pedigrees are available as File S2 and the SNP genotypes as File S3 based on the chosen SNP markers are provided in File S4.

As in Legarra *et al.* (2008), we also added the random effect of cage in the WGP model of [2]; *i.e.*, $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{g} + \mathbf{W}\mathbf{u} + \mathbf{S}\mathbf{c} + \mathbf{e}$, where $\mathbf{c} \sim N(0, \mathbf{I}\sigma_c^2)$ and \mathbf{S} is the cor-

responding incidence matrix with all other terms defined as before. Furthermore, we specified Gelman's prior $\sigma_c^2 \sim \chi^{-2}(-1, 0)$ on σ_c^2 in addition to all previously provided prior specifications. Also, as per Legarra *et al.* (2008), we chose to use the data provided on body weight at 6 weeks that was already precorrected for fixed effects such that \mathbf{X} was a column vector of ones and β consisted of just an overall mean. Missing SNP genotypes were simply imputed from binary distributions on the basis of their corresponding allelic frequencies in the data set following Legarra *et al.* (2008). We adopted the same within-family cross-validation technique as described in Legarra *et al.* (2008) by randomly partitioning each family into two. This partitioning was replicated 20 times to obtain 20 different nearly equal-sized partitions of training and validation data subsets. Also, as in Legarra *et al.* (2008), we compared the various methods using predictive abilities, defined as the correlation between phenotypes in the validation subset and their corresponding predictions on the basis of their inferences from the training data subset.

Application on simulated genomic data from Hickey and Gorjanc

To provide a benchmark comparison of our proposed methods with competing methods in other articles in this issue, we analyze simulated data sets provided by and described in detail by Hickey and Gorjanc (2012). They generated 10 replicated data sets for each of four different traits whereby 9000 QTL effects were generated for trait 1 and 900 QTL effects were generated for trait 2. Traits 3 and 4 mirrored traits 1 and 2, respectively, with the further requirement that the MAF for these QTL was <0.30 . Since we were permitted to simultaneously run 144 jobs on the high-performance computing cluster at MSU (hpcc.msu.edu), we chose to compare the four methods for each of the four traits on each of the first nine data sets ($4 \times 4 \times 9 = 144$). For all analyses, training data were based on 2000 animals in generations 4 and 5 whereas TBV were provided on 500 animals within generations 6, 8, and 10. To facilitate computing tractability, we saved every tenth SNP marker that had a MAF >0.20 . This led to a range of 2884–2952 SNP markers and an average LD between adjacent markers from 0.16 to 0.17 across the nine replicates. The actual SNP markers chosen from each of the nine data sets are provided in File S5 based on the original data (File S6) as provided by Hickey and Gorjanc (2012). All four models also included polygenic effects. Antedependence methods were directly compared with their classical counterparts for accuracy (correlation of EBV with TBV) and bias (deviation of slope from 1 from regressing TBV on EBV) in these latter validation generations using a Wilcoxon signed rank test.

Bayesian inference

For each of the four methods, BayesA, BayesB, ante-BayesA, and ante-BayesB in both our simulation study and the heterogeneous stock mice application, we ran MCMC for 50,000 cycles of burn-in followed by an additional 300,000 cycles; for the benchmark data from Hickey and Gorjanc

Table 1 Summary statistics for six different marker densities in the simulation study over 20 replicates

Marker density level ^a	Average no. of markers per replicate	Average distance between adjacent marker loci (cM) per replicate	Average r^2 between adjacent marker loci per replicate
1	108	0.918	0.15
2	145	0.689	0.18
3	217	0.459	0.21
4	311	0.321	0.24
5	545	0.184	0.27
6	2182	0.046	0.31

^a Marker density levels 1–6 pertain to saving every 20th, 15th, 10th, 7th, and 4th and every single SNP marker from a single 1-M chromosome within each data replicate.

(2012), the corresponding numbers were 80,000 and 1,000,000, respectively. Every tenth MCMC cycle was subsequently saved for inference post burn-in. We monitored MCMC convergence via inspection of trace plots and determined the effective sample size (ESS) for number of random draws from the joint posterior density for all key hyperparameters using the R package CODA (Plummer *et al.* 2006). The larger number of MCMC cycles for the Hickey–Gorjanc data were based on ensuring that ESS for all hyperparameters at least exceeded 100. Inferences were primarily based on the posterior means and posterior standard deviations for key parameters, including those hyperparameters that characterize genetic architecture.

Prior specifications

For all analyses in this article, we chose $\alpha_\pi = 10$ and $\beta_\pi = 1$ in both BayesB and ante-BayesB to reflect the prior belief that most of the markers will not be associated with any genetic effects; however, the dispersion of this corresponding beta distribution is still large enough such that values of π_g (π_δ) close to 0.70 are plausible. On the basis of preliminary runs, we also found that this prior specification led to superior mixing properties of the MCMC chains over a naïve Uniform(0,1) prior, yet facilitated domination of data over prior information since $\alpha_\pi + \beta_\pi \ll m$.

For BayesB and ante-BayesB, we always specified $\alpha_s = \beta_s = 0.1$ for the Gamma prior on s_g^2 (s_δ^2). For the ante-dependence-based models, we specified $\mu_{t0} = 0$, $s_{t0}^2 = 1$, $\nu_t = -1$, and $s_t^2 = 0$, *i.e.*, a standard normal prior on μ_t and Gelman’s prior on σ_t^2 . We also always specified a flat prior on β by defining $\mathbf{V}_\beta^{-1} = \mathbf{0}$. Prior specifications for all other parameters (*e.g.*, variance components) were based on those previously recommended in this article.

Results

Simulation study

For the six different marker densities, the average distances between adjacent markers ranged from 0.046 to 0.918 cM over the 20 replicates whereas the average LD between adjacent markers, measured by r^2 values, ranged from 0.15 to 0.31, as shown in Table 1. Among the 30 chosen QTL within

each of the 20 replicates, anywhere between 6 and 11 of the QTL had variances $>2\%$ of the total genetic variance.

Inferences on key hyperparameters

It is important to recognize that none of the modeling assumptions behind BayesA, BayesB, anteBayesA, or ante-BayesB truly match the data generation model based on thousands of generations of LD created between markers and QTL, even for simulated data. This goes beyond the fact that the QTL effects were drawn from reflected Gamma distributions in our simulation study as typically done (*e.g.*, Meuwissen *et al.* 2001; Meuwissen and Goddard 2010). That is, the process of recombination over thousands of generations in terms of how it generates LD between QTL and SNP markers is not explicitly captured in any known WGP model, including any of the competing models, especially when the effects of neighboring SNP markers rather than the causal QTL effects are being estimated. Hence, there is no way to surmise the “true” values of key hyperparameters, whether for s_g^2 , ν_g , or π_g in BayesA or BayesB or for s_δ^2 , ν_δ , π_δ , μ_t , or σ_t^2 in anteBayesA or anteBayesB. However, one should anticipate that estimates of s_g^2 or s_δ^2 should be inversely related to marker density, since they closely represent the mean value of the variance components $\{\sigma_{g_j}^2\}_{j=1}^m$ or $\{\sigma_{\delta_j}^2\}_{j=1}^m$, respectively, accounted for by each SNP. Indeed, we observe this phenomenon in the comparison between BayesA and anteBayesA in Figure 1A. We also note a similar comparison between s_δ^2 and s_g^2 for BayesB vs. anteBayesB in Figure 1B, but further recognize that the corresponding estimates of s_g^2 and s_δ^2 are roughly 1 order of magnitude greater than those seen in Figure 1A. That is, s_g^2 and s_δ^2 specify a typical value for $\sigma_{g_i}^2$ and $\sigma_{\delta_i}^2$, respectively, over many more loci in (ante)BayesA than their (ante)BayesB counterparts. In spite of the lower values observed in Figure 1A, however, there was a significant difference ($P < 0.01$) between s_g^2 and s_δ^2 when $r^2 \geq 0.21$.

As marker density increased, we also expected that the estimates of π_g or π_δ should increase as well; that is, it becomes increasingly unlikely that individual SNP markers become associated with a particular QTL with greater marker density. Indeed we observed this in Figure 2. It was particularly interesting that the posterior means of π_δ were generally lower than that of π_g , with differences widening with increasing marker density (*i.e.*, LD level) such

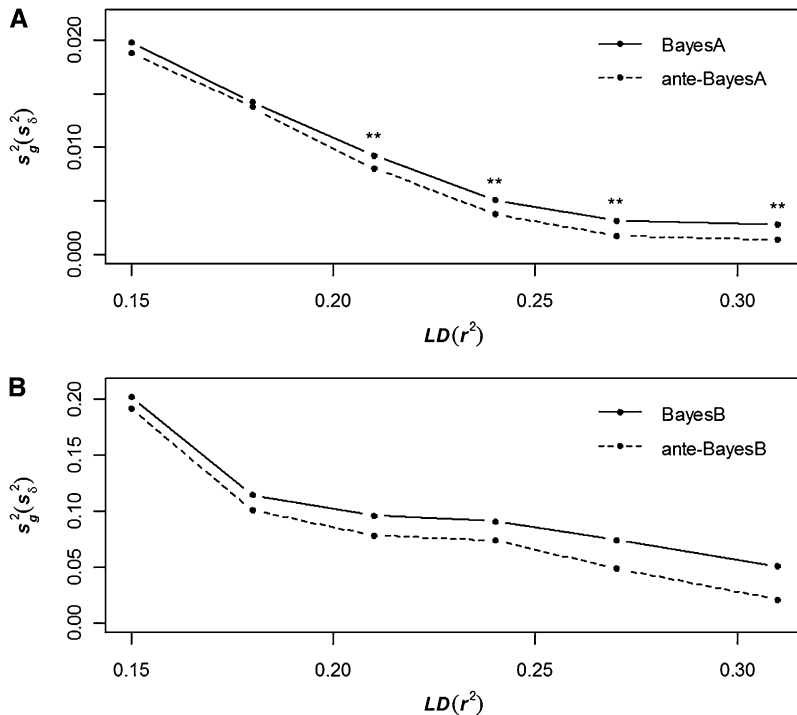


Figure 1 Average posterior means of s_g^2 (BayesA, BayesB) and s_δ^2 (ante-BayesA, ante-BayesB) across 20 replicates for six different levels of LD comparing BayesA and ante-BayesA (A) and BayesB vs. ante-BayesB (B). Significant differences in posterior means between competing methods at each LD level are indicated: (***) $P < 0.001$.

that the differences were significant beyond $r^2 = 0.24$ ($P < 0.01$). Note the subtle difference in interpretation between π_g and π_δ as π_g pertains to the probability of nonassociation for the corresponding SNP whereas π_δ pertains to the probability of nonassociation conditional on a neighboring SNP.

The estimates of ν_g and ν_δ also changed as a function of marker density for anteBayesA vs. BayesA in Figure 3A and for anteBayesB vs. BayesB in Figure 3B. Specifically, the posterior means of ν_g and particularly of ν_δ both decrease with increasing marker intensity. Since these parameters, respectively, characterize the heterogeneity of $\sigma_{g_j}^2$ and $\sigma_{\delta_j}^2$ across SNP or, alternatively, the heaviness of the tails for the resulting marginal Student t distribution on g_j and δ_j across SNP, our results imply that these hierarchical methods, and particularly those based on nonstationary first-order antedependence correlation structures, identify SNP with large effects as being more outlying relative to a normal distribution when marker density increases. However, these differences between ν_g and ν_δ were not seen to be statistically significant at any marker density.

Figures S1 and S2 in supporting information File S1 shows the average posterior means for μ_t and σ_t^2 against LD level across the 20 replicates under both ante-BayesA and ante-BayesB. There was no evidence ($P > 0.01$) across these 20 replicates that the posterior means of μ_t were different from zero at any LD level; however, at higher LD levels, the posterior means tended to converge to zero as anticipated. Similarly, the posterior estimates for σ_t^2 were also lower at higher LD levels. Again, this was somewhat anticipated since there should be less disparity in different values of the antedependence parameters ($t_{j,j-1}$) between adjacent markers with increasing marker intensity.

Comparison between true and estimated total breeding values

The average accuracies of the EBV over the 20 replicated data sets are plotted as a function of the average r^2 (i.e., the different marker densities) between adjacent markers for the four different methods in Figure 4. As anticipated, given the simulated genetic architecture of few QTL, the accuracies for the BayesB methods were consistently greater than

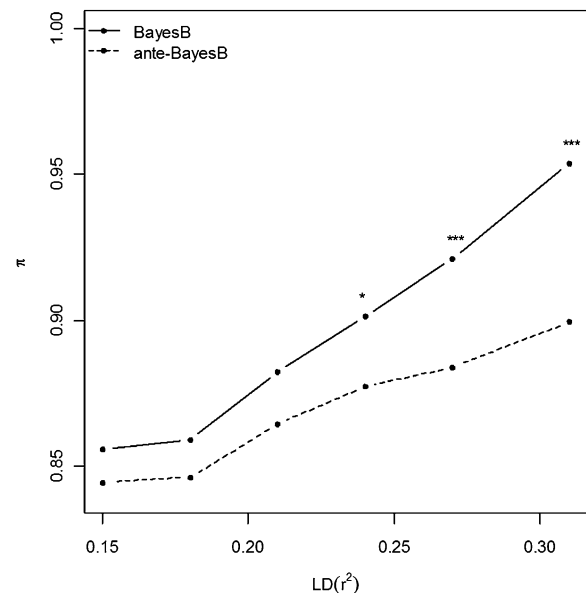


Figure 2 Average posterior means of π_g (BayesB) vs. π_δ (ante-BayesB) across 20 replicates as a function of six different LD levels. Significant differences in posterior means between competing methods at each LD level are indicated by (*) $P < 0.01$, or (***) $P < 0.0001$.

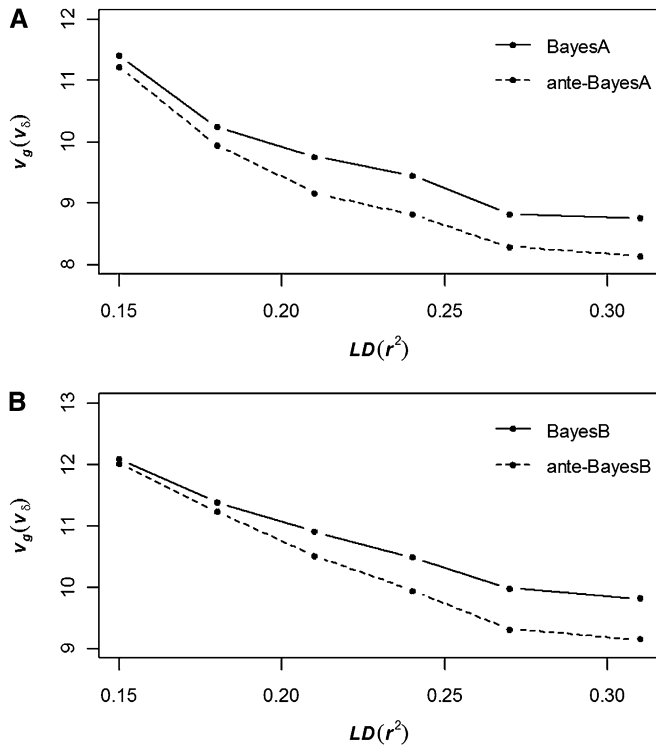


Figure 3 Average posterior means of v_g (BayesA, BayesB) and v_δ (ante-BayesA, ante-BayesB) across 20 replicates for six different levels of LD comparing BayesA and ante-BayesA (A) and BayesB vs. ante-BayesB (B). No significant differences ($P > 0.01$) were determined between the two sets of competing procedures at each LD level.

their corresponding BayesA counterparts at all marker densities. Also, ante-BayesA and ante-BayesB outperformed their classical counterparts with differences increasing with LD level. Specifically, ante-BayesA had significantly greater accuracies compared to conventional BayesA, as did ante-BayesB compared to BayesB ($P < 0.01$), when average LD levels exceeded $r^2 = 0.24$.

We anticipated that the antedependence parameters $t_{j,j-1}$'s would have greater importance at higher marker densities. To demonstrate this, we standardized the posterior means of these parameters as a ratio over their posterior standard deviations, i.e., $\tilde{t}_{j,j-1} = E(t_{j,j-1} | y) / \sqrt{\text{var}(t_{j,j-1} | y)}$, for each analysis. We then determined the proportion of these $\tilde{t}_{j,j-1}$ whose absolute value exceeded an arbitrary value of 2 for each data replicate and marker density analysis to indicate the relative importance of these antedependence parameters. We present boxplots of these proportions across the 20 replicates for ante-BayesA and for ante-BayesB in Figure S3 in File S1. We anticipated and noted that a higher proportion of $\tilde{t}_{j,j-1}$ exceeded 2 in data sets characterized by higher marker densities, thereby indicating that, in general, nonstationary serial correlation between adjacent markers becomes increasingly more important with higher levels of LD. We believe this phenomenon is responsible for driving the differences in accuracies between ante-BayesA (ante-BayesB) and BayesA (BayesB) with increasing LD levels as seen earlier in Figure 4.

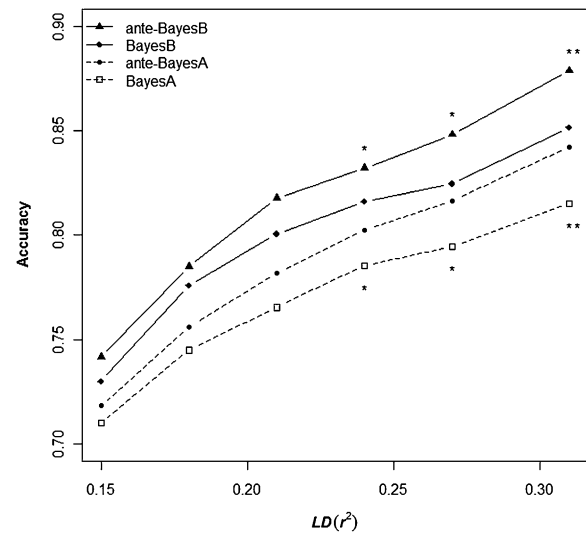


Figure 4 Average accuracies of estimated breeding value across 20 replicates for analyses based on each of six LD levels. Differences in accuracy between BayesA and ante-BayesA (circle and square) and between BayesB with ante-BayesB (triangle and diamond) indicated as significant by (*) $P < 0.01$ or (**) $P < 0.001$.

Potential merit for GWAS

Hierarchical methods that are similar to BayesB, in that they jointly infer upon all SNP effects, have been increasingly advocated as tools for GWAS (Hoggart *et al.* 2008; Lee *et al.* 2008; Logsdon *et al.* 2010). Figure S4 in File S1 shows the average (across 20 replicates) posterior mean probabilities of identifying the largest QTL by genetic variance within each replicate using BayesB and ante-BayesB, respectively. These estimated posterior probabilities increased with LD level for both models but were significantly greater for ante-BayesB than for BayesB with statistical significance also increasing with LD or marker density. That is, the precision for detecting QTL was increasingly greater for ante-BayesB than for BayesB at higher LD levels. We observed this consistently across data replicates with the ability of ante-BayesB to better track causal variants increasing with marker density (Figure S5 in File S1).

Application to heterogeneous stock mice data

We summarize posterior inferences of key parameters using BayesA and BayesB in Table S1 and for their antedependence counterparts in Table S2 (see File S1) for the heterogeneous stock mice data. Inferences on σ_u^2 , σ_c^2 , and σ_e^2 were consistent with results previously reported by Legarra *et al.* (2008). As expected from our simulation study, the estimates for v_g (v_δ) and s_g^2 (s_δ^2) were substantially greater for BayesB (ante-BayesB) than for BayesA (ante-BayesA). Although the posterior mean for π_g of 0.81 (BayesB) was only slightly larger than $\pi_\delta = 0.80$ (ante-BayesB), the posterior mean of s_δ^2 was substantially larger in ante-BayesB than s_g^2 in BayesB. The average estimates \pm empirical standard errors of predictive ability correlations over the 10 cross-validation partitions of training and validation data subsets were 0.57 ± 0.01 , 0.62 ± 0.01 , 0.60 ± 0.01 , and 0.66 ± 0.01 for BayesA, BayesB, ante-BayesA, and

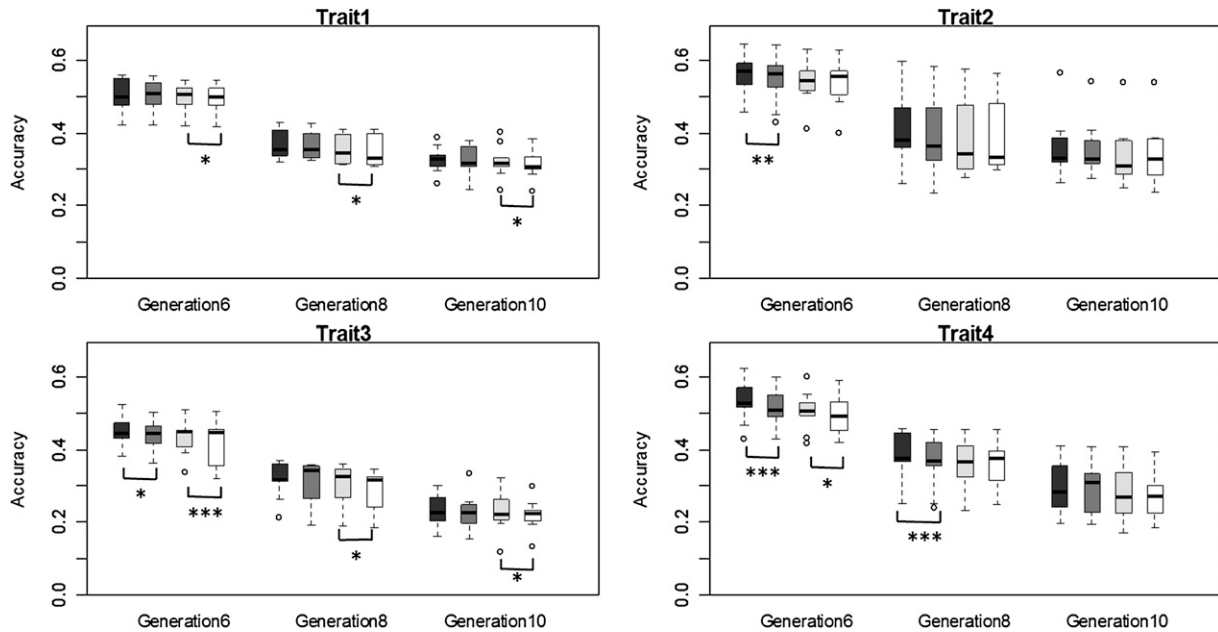


Figure 5 Boxplots of average accuracies of estimated breeding value across nine replicates for four traits in generations 6, 8, and 10 for benchmark data from Hickey and Gorjanc (2012). Differences in accuracy between anteBayesB (solid bars) and BayesB (dark shading) and between anteBayesA (light shading) with BayesA (open bars) indicated as significant by (*) $0.05 < P < 0.10$, (**) $0.01 < P < 0.05$, or (***) $P < 0.01$.

ante-BayesB, respectively. The differences between BayesA with ante-BayesA and BayesB with ante-BayesB were both determined to be statistically significant ($P < 0.005$), indicating the relative advantage of the antedependence methods. Furthermore, BayesB and ante-BayesB had significantly greater predictive abilities than BayesA and ante-BayesA, respectively ($P < 0.001$).

Application to Hickey and Gorjanc data

Average posterior means for key hyperparameters for each of the four methods across the nine replicates are provided in Table S3 in File S1, whereas the corresponding average ESS are provided in Table S4 in File S1. Estimates of π_g (π_δ) and s_g^2 (s_δ^2) were lower whereas estimates of ν_g (ν_δ) are higher for traits with higher numbers of QTL (traits 1 and 3) compared to those with lower numbers of QTL (traits 2 and 4), relative to the same number of markers.

A side-by-side comparison of the accuracies of the four methods across the validation generations (6, 8, and 10) is provided in Figure 5. It is remarkable to note that ante-BayesA had generally significantly greater accuracies than BayesA for traits 1 and 3 (larger numbers of QTL) that was still maintained until generation 10, whereas ante-BayesB had generally significantly greater accuracies than BayesB for traits 2 and 4 (lower numbers of QTL) but only in generations 6 and 8. An assessment of bias of the four procedures based on regressing TBV on EBV is provided in Figure S6 in File S1. For all traits, all four methods had some significant bias in generation 6 but not in later generations.

Discussion

In this article, we extend two very popular Bayesian methods, BayesA and BayesB, for WGP to model potential nonstationary correlations between SNP effects in close proximity to QTL. We demonstrated using a small-scale simulation study that the accuracies of our proposed first-order antedependence extensions, labeled ante-BayesA and ante-BayesB, were greater than their classical counterparts with differences increasing with marker density. This result was anticipated given that the magnitude and importance of the antedependence parameters $\{t_{j-1}\}_{j=2}^m$ in \mathbf{T} should increase as marker densities increase. To further illustrate the importance of modeling nonstationary correlations, rather than basing correlations between SNP effects purely as a function of distance (Gianola *et al.* 2003), we observed the magnitude of the posterior means of $\{t_{j-1}\}_{j=2}^m$ at the various locations through the chromosome for the first four replicates using ante-BayesA and ante-BayesB (see Figure S7 in File S1). These posterior means tended to be rather large in absolute value in the general vicinity of the major QTL. This result was anticipated since each QTL is likely tracked by several SNP, each in incomplete LD with the QTL (Goddard and Hayes 2009). Interestingly, there appeared to be a greater spread in these posterior means around a greater number of QTL using ante-BayesB compared to ante-BayesA.

We realize that the order in which Equation 3 is specified for the antedependence methods is rather arbitrary; *i.e.*, one might specify Equation 3 from the end of the p -arm to the end of the q -arm of a chromosome or vice versa. For instance, instead of specifying Equation 3 from $j = 1, 2, \dots, m$, we

might have also modeled antedependence in the opposite direction, *i.e.*, from $j = m, m-1, \dots, 1$. It has been demonstrated by Zimmerman and Nunez-Anton (2010) in the context of longitudinal data analysis that the (co)variances based on a first-order antedependence model are invariant to directionality as long as the relative order is correctly specified. To illustrate this, we reanalyzed the first four replicates on the basis of the highest average marker density ($r^2 = 0.31$), again using ante-BayesA and ante-BayesB but this time specifying Equation 3 in the opposite direction from what was used previously. We plotted the posterior means ($\bar{\mathbf{g}}$) of \mathbf{g} on the basis of the analysis in the original direction against the same estimates on the basis of the analysis in the opposite direction (Figure S8 in File S1), further demonstrating that inferences on EBV are invariant to direction. Noting that the EBV are linear functions of $\bar{\mathbf{g}}$, *i.e.*, $\mathbf{Z}\bar{\mathbf{g}}$, we noted even greater consistency for EBV between the two directions for these same four replicates in Figure S9 in File S1.

Given that the accuracy of EBV was greater using the antedependence-based procedures compared to their classical counterparts, we examined the elements of $\bar{\mathbf{g}}$ for each marker between the two different classes of models within each of the first four replicates and the highest average marker density (Figure S10 in File S1). It was interesting to note that these elements were more shrunk to zero using the antedependence-based procedures than were the conventional counterparts. Given the specification of nonstationary correlations between effects of adjacent SNP markers, the effective number of SNP may be considered to be somewhat lower using the antedependence-based procedures, particularly in regions containing a major QTL. That is, larger elements of $\bar{\mathbf{g}}$ using the antedependence-based models will tend to be more highly correlated and hence shrunk closer to zero than elements of $\bar{\mathbf{g}}$ derived from their conventional counterparts. Goddard *et al.* (2009) have previously described the optimal statistical properties of $\bar{\mathbf{g}}$ when \mathbf{g} is specified as a set of “random effects” having an exchangeable distribution, such as the normal or Student’s *t*. However, these properties are better realized when a more appropriate correlation structure is specified, as would be true, for example, modeling polygenic effects with a classical animal model (Henderson 1984). The more optimal properties of $\bar{\mathbf{g}}$ using the antedependence-based models was also partly reflected in the earlier EBV comparisons with their classical counterparts.

In spite of the limited scale of our simulation studies, it has been demonstrated that inferences on accuracy based on 100 individuals and a genome length of 1 M is roughly equivalent to inferences derived from 3000 individuals and a genome length of 30 M (Meuwissen and Goddard 2010), the latter of which might depict a more common scenario in livestock populations. We also based all of our simulation work on a heritability of 50%; for situations with lower or higher heritabilities, we would naturally expect the accuracies to, respectively, decrease or increase accordingly in concert with previous simulation results (Calus and

Veerkamp 2007) and/or analytical derivations (Meuwissen and Goddard 2010); however, we believe that there is no reason to believe that the antedependence methods would not outperform the conventional Bayesian WGP methods in these situations as well. This was further substantiated by our analyses of the data from Hickey and Gorjanc (2012), which were based on a heritability of 25%. In our own simulation study, there were around 2200 markers per the single 1-M chromosome using the highest average LD level of $r^2 = 0.31$, whereas there were around 100 markers per the single chromosome with the lowest average LD level of $r^2 = 0.15$. Using Meuwissen and Goddard (2010), these two specifications are, respectively, analogous to a panel of 60,000 SNP markers and to a panel of 3000 SNP markers for a 30-M genome; commercially developed panels having roughly these same numbers of SNP markers are now widely available for cattle (Wiggans *et al.* 2011). On the basis of the results of our work, we anticipate that the antedependence-based methods, compared to their classical counterparts, would lead to even greater accuracies with higher density SNP marker panels ($m > 500,000$) that are being developed for livestock or for situations in which there is sequence data (Meuwissen and Goddard 2010). Along those lines, we anticipate that these methods would also perform better in populations where LD is greater between markers due to other phenomena, *e.g.*, selection history.

Our simulation studies were also based on a particular genetic architecture; *i.e.*, 30 QTL that were randomly distributed throughout a 1-M chromosome (or equivalently, 900 QTL for a 30-M genome). Although this is not the focus of our article, we realize that genetic architecture (*i.e.*, number of QTL, average QTL substitution effect, marker density, etc.) can affect the relative merit of BayesA, BayesB, and GBLUP on the basis of other studies in which key hyperparameters such as π_g , ν_g , and s_g^2 are arbitrarily specified to be known (Daetwyler *et al.* 2010; Meuwissen and Goddard 2010). That is, the greater the number of QTL, each with small effects, relative to the number of SNP markers, the more likely the genetic architecture reflects the GBLUP assumptions ($\pi_g = 1$, $\nu_g \rightarrow \infty$ such that $\sigma_{g_j}^2 = s_g^2 \forall j$). Conversely, BayesB would be favored in the situation in which SNP marker density is high relative to the number of QTL ($\pi_g < 1$). However, we believe that formal comparisons in data fit between BayesA, BayesB, and GBLUP, along with ante-BayesA and ante-BayesB, are not entirely necessary since ante-BayesB represents the most general model. As previously noted, ante-BayesA is a special case of BayesA, as is ante-BayesB of BayesB, when $\mathbf{T} = \mathbf{0}$ such that then $\pi_\delta = \pi_g$, $\nu_\delta = \nu_g$, and $s_\delta^2 = s_g^2$. Furthermore, BayesB becomes BayesA as $\pi_g \rightarrow 1$, whereas BayesA becomes GBLUP as $\nu_g \rightarrow \infty$. Nevertheless, our claim that one needs only to fit ante-BayesB, rather than any of the other three competing submodels, vitally depends upon reliable inferences being provided on these key hyperparameters defining genetic architecture, rather than arbitrarily specifying them (Daetwyler *et al.* 2010; Meuwissen and Goddard 2010) or estimating a subset thereof (Habier

et al. 2011). We provide details on MCMC inference strategies on these and other unknown parameters in File S1. We are currently pursuing more suitable inferential strategies for variable selection (O'Hara and Sillanpaa 2009) when inferring upon π_g or π_δ . Also, although our proposed antedependence methods seem to work well under additive genetic model assumptions, it is not clear how well they may perform in the presence, for example, of extensive nonadditive gene action where nonparametric approaches may be warranted (Gianola *et al.* 2010). Nevertheless, even in the extensive presence of such phenomena, genetic variance is still considered to be primarily additive (Hill *et al.* 2008).

Although the scope of this work was focused on the potential merit of these antedependence models for WGP, we suggested earlier that there may be also merit in using these models in GWAS in both livestock and human populations. It has become increasingly recognized that GWAS procedures based on joint analyses of all SNP markers are more powerful than the conventional series of single SNP analyses. Our results suggest that modeling nonstationary correlations between SNP effects will further augment this power. At any rate, we recognize that for reasonably accurate GWAS, a greater marker density (m) per chromosome and sample size (n) should be considered (e.g., Meuwissen and Goddard 2010) than those studied in this article; *i.e.*, most of the posterior probabilities reported in Figures S4 and S5 in File S1 are too low to be of practical benefit in current applications.

We also acknowledge that our ante-BayesA and ante-BayesB models increase the computational load relative to their conventional counterparts. Since m is typically large, the computing time for the proposed antedependence models is bottlenecked primarily by the m elements of δ , the m diagonal elements of Δ , and the $m - 1$ nonzero elements of T . Similarly, computing time for the two conventional methods, BayesA and BayesB, is primarily restricted by the dimension of δ and Δ , *i.e.*, roughly two-thirds as many variables for the antedependence-based models, ignoring the remaining parameters such as variance components and hyperparameters. Hence, the computing time for the antedependence-based procedures should be somewhat less than one-third greater than for their conventional counterparts. Indeed, we discovered from our simulation study that computing time for all four competing models were linear in m with the antedependence-based models taking <30% greater computing time than the conventional counterparts for the wide range of values of m considered in this article. We recognize for much larger number of SNP markers than those pursued in this study that alternative algorithmic adaptations already developed for models similar to conventional BayesA or BayesB, such as those based on the EM algorithm (Shepherd *et al.* 2010) or variational Bayes (Logsdon *et al.* 2010), would be worth exploring.

We believe the proposed antedependence models provide opportunities for further study and extension. For example, it has been previously recognized that basing inferences on allelic effects on the use of multiple marker haplotypes rather

than single markers increases accuracy of WGP (Calus *et al.* 2008; Villumsen *et al.* 2008) or GWAS (Grapes *et al.* 2004). Given the difficulty in how to appropriately specify these haplotypes, we believe our antedependence-based methods may help bridge these two different strategies as the effects of adjacent SNP markers connected by large values of $t_{j,j-1}$ may somewhat determine "effective haplotype" effects. We also think that our antedependence specifications might facilitate multiple breed inference if, for example, genomic effect differences between breeds is primarily due to differences in SNP associations with QTL, as partly manifested in T.

Acknowledgments

This project was supported by Agriculture and Food Research Initiative Competitive Grant no. 2011-67015-30338 from the U.S. Department of Agriculture National Institute of Food and Agriculture. Additional support from the Quantitative Biology Program, the Department of Animal Science, and AgBioResearch at Michigan State University is also gratefully acknowledged.

Literature Cited

- Abasht, B., E. Sandford, J. Arango, P. Settar, J. E. Fulton *et al.*, 2009 Extent and consistency of linkage disequilibrium and identification of DNA markers for production and egg quality traits in commercial layer chicken populations. *BMC Genomics* 10: S2.
- Bello, N. M., J. P. Steibel, and R. J. Tempelman, 2010 Hierarchical Bayesian modeling of random and residual variance-covariance matrices in mixed effects models. *Biom. J.* 52: 297–313.
- Bohmanova, J., M. Sargolzaei, and F. S. Schenkel, 2010 Characteristics of linkage disequilibrium in North American Holsteins. *BMC Genomics* 11: 421.
- Calus, M. P. L., and R. F. Veerkamp, 2007 Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *J. Anim. Breed. Genet.* 124: 362–368.
- Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp, 2008 Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 178: 553–561.
- Coster, A., J. W. M. Bastiaansen, M. P. L. Calus, J. A. M. van Arendonk, and H. Bovenhuis, 2010 Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. *Genet. Sel. Evol.* 42: 9.
- Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams, 2010 The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185: 1021–1031.
- Daniels, M. J., and M. Pourahmadi, 2002 Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* 89: 553–566.
- de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra *et al.*, 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182: 375–385.
- de los Campos, G., D. Gianola, and D. B. Allison, 2010 Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat. Rev. Genet.* 11: 880–886.
- de Roos, A. P. W., B. J. Hayes, R. Spelman, and M. E. Goddard, 2008 Linkage disequilibrium and persistence of phase in Holstein, Friesian, Jersey and Angus cattle. *Genetics* 179: 1503–1512.

- Du, F. X., A. C. Clutter, and M. M. Lohuis, 2007 Characterizing linkage disequilibrium in pig populations. *Int. J. Biol. Sci.* 3: 166–178.
- Gelman, A., 2006 Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.* 1: 515–533.
- Gianola, D., M. Perez-Enciso, and M. A. Toro, 2003 On marker-assisted prediction of genetic value: beyond the ridge. *Genetics* 163: 347–365.
- Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando, 2009 Additive genetic variability and the Bayesian alphabet. *Genetics* 183: 347–363.
- Gianola, D., X.-L. Wu, E. Manfredi, and H. Simianer, 2010 A non-parametric mixture model for genome-enabled prediction of genetic value for a quantitative trait. *Genetica* 138: 959–977.
- Goddard, M. E., and B. J. Hayes, 2009 Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.* 10: 381–391.
- Goddard, M. E., N. R. Wray, K. Verbyla, and P.M. Visscher, 2009 Estimating effects and making predictions from genome-wide marker data. *Stat. Sci.* 24: 517–529.
- Grapes, L., J. C. M. Dekkers, M. F. Rothschild, and R. L. Fernando, 2004 Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. *Genetics* 166: 1561–1570.
- Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* 12: 186.
- Hayes, B. J., P. J. Bowman, A.J. Chamberlain, and M. E. Goddard, 2009 Invited review: genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92: 433–443.
- Henderson, C. R., 1976 A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* 32: 69–83.
- Henderson, C. R., 1984 *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph, Canada.
- Hickey, J. M., and G. Gorjanc, 2012 Simulated data for genomic selection and GWAS using a combination of coalescent and gene drop methods. *G3: Genes, Genomes, Genetics* 2: 425–427.
- Hill, W. G., M. E. Goddard, and P. M. Visscher, 2008 Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* 4: e1000008.
- Hoggart, C. J., J. C. Whittaker, M. De Iorio, and D. J. Balding, 2008 Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. *PLoS Genet.* 4: e1000130.
- Kizilkaya, K., and R. J. Tempelman, 2005 A general approach to mixed effects modeling of residual variances in generalized linear mixed models. *Genet. Sel. Evol.* 37: 31–56.
- Lee, S. H., J. H. van der Werf, B. J. Hayes, M. E. Goddard, and P. M. Visscher, 2008 Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet.* 4: e1000231.
- Legarra, A., C. Robert-Granie, E. Manfredi, and J.-M. Elsen, 2008 Performance of genomic selection in mice. *Genetics* 180: 611–618.
- Logsdon, B., G. Hoffman, and J. Mezey, 2010 A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics* 11: 58.
- Lorenz, A. J., S. M. Chao, F. G. Asoro, E. L. Heffner, T. Hayashi *et al.*, 2011 Genomic selection in plant breeding: knowledge and prospects, pp. 77–123 in *Advances in Agronomy*, vol. 110. Elsevier–Academic Press, San Diego.
- Meuwissen, T., and M. Goddard, 2010 Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* 185: 623–631.
- Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157: 1819–1829.
- O’Hara, R. B., and M. J. Sillanpaa, 2009 A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal.* 4: 85–117.
- Plummer, M., N. Best, K. Cowles, and K. Vines, 2006 CODA: convergence diagnostics and output analysis for MCMC. *R News* 6: 7–11.
- Shepherd, R. K., T. H. Meuwissen, and J. A. Woolliams, 2010 Genomic selection and complex trait prediction using a fast EM algorithm applied to genome-wide markers. *BMC Bioinformatics* 11: 529.
- Sorensen, D., and D. Gianola, 2002 *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*, Springer-Verlag, New York.
- Valdar, W., L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman *et al.*, 2006a Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.* 38: 879–887.
- Valdar, W., L. C. Solberg, D. Gauguier, W. O. Cookson, J. N. P. Rawlins *et al.*, 2006b Genetic and environmental effects on complex traits in mice. *Genetics* 174: 959–984.
- VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91: 4414–4423.
- Villumsen, T. M., L. Janss, and M. S. Lund, 2008 The importance of haplotype length and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet.* 126: 3–13.
- Wiggans, G. R., P. M. VanRaden, and T. A. Cooper, 2011 The genomic evaluation system in the United States: past, present, future. *J. Dairy Sci.* 94: 3202–3211.
- Yi, N., and S. Xu, 2008 Bayesian Lasso for quantitative trait loci mapping. *Genetics* 179: 1045–1055.
- Zimmerman, D. L., and V. A. Nunez-Anton, 2010 *Ante-dependence Models for Longitudinal Data*. Chapman & Hall, London/New York.

Edited by Dirk-Jan de Koning, Lauren M. McIntyre, and Matias Kirst